Total Error in a Plug-in Estimator of Level Sets

Amparo Baillo Moreno*

Abstract

Given a probability density *f* on *R^d*, the minimum volume set of probability content á can be estimated by the level set of the same probability content corresponding to a kernel estimator of *f*. We obtain convergence rates for this plug-in estimator with respect to a measure-based distance between sets. This distance has a straightforward interpretation in the context of cluster analysis.

Keywords: Empirical process theory; kernel density estimates; level set estimation; rates of convergence.

*Baillo, Statistics and Econometrics Department, University Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), e-mail: abaillo@est-econ.uc3m.es. I would like to thank Antonio Cuevas for his careful reading of the manuscript and his suggestions, which led to improvement of the results. A large part of this work was developed at the Statistics Semester that took place in the Institut Henri Poincare (Paris) in 2001, and was supported by the CNRS.

1. Introduction

1.1 Motivation of the problem

Given a probability measure F on $\mathbb{R}^d$ with density f, we are interested in nonparametric estimation of the level sets $\{f > c\}$, for small $c > 0$. The study of these regions with high mass concentration is useful, for example, in cluster analysis. Hartigan (1975) defined clusters in a population as the connected components of the level set $\{f > c\}$. This means that observations falling outside this set will remain unclassified.

Density level sets may also be employed to develop a quality control scheme. Following the lines of the nonparametric set-based proposal in Devroye and Wise (1980), we will decide that a manufacturing process is out of control if a new observation belongs to the set $\{f \leq c\}$. In both frameworks it makes sense to estimate $\{f > c\}$ by $\{f_n > c\}$, where f_n is a nonparametric estimator of f (see Cuevas, Febrero and Fraiman 2000).

We will focus on a particular type of level sets: minimum volume sets $\{f > c_\alpha\}$ with probability content α (for $0 < \alpha < 1$), which have been thoroughly studied in the context of robust statistics. For instance, they have been used to construct robust estimators of location and dispersion (see Polonik 1997 and references therein). On the other hand, in the nonparametric detection procedure mentioned above the estimation of these minimum volume sets would arise from the wish to bound (by $1 - \alpha$) the probability of giving a false alarm $F\{f_n \leq c\}$.

1.2 Statement of the problem. Notation

Let f be the unknown density function of a probability distribution F on $\mathbb{R}^d$. Take a random sample $X_1, \ldots, X_n$ of independent observations from f. Let f_n be the kernel

estimator of f with kernel K and smoothing parameter $h = h_n$

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)$$

where K is a probability density, $K_h(x) := h^{-d}K(x/h)$ and

$$h_n \to 0, \quad \frac{nh_n^d}{\log n} \to \infty \quad \text{as} \quad n \to \infty. \tag{1}$$

We will use the following notation

$$F(A) = \int_A f, \qquad \tilde{F}_n(A) = \int_A f_n, \qquad \text{for any } A \in \mathcal{B}_{\mathbb{R}^d}.$$

Consider a fixed $\alpha \in (0,1)$. Let the quantile function based on F and Leb, the Lebesgue measure, be defined as

$$V(\alpha) = \text{Leb}\{f > c_\alpha\} \quad \text{where } c_\alpha = \sup\{c : \int_{\{f>c\}} f > \alpha\}.$$

It can be seen that, if f is bounded and $\text{Leb}\{f = c_\alpha\} = 0$, then $\int_{\{f>c\}} f$ is continuous at $c = c_\alpha$, V is continuous at α and $F\{f > c_\alpha\} = \alpha$. The statement $\text{Leb}\{f = c\} = 0$ just means that f is not flat at level c, an unavoidable condition in nonparametric estimation of level sets $\{f > c\}$ (see, for example, Tsybakov 1997, Polonik 1995, 1997). In fact, from now on we will use the following assumption to prevent the existence of flat parts in a neighbourhood of c_α.

(F1) *There exist constants $a, b, \gamma, C > 0$ such that, for all $c \in [a, b]$ and $\epsilon > 0$ small enough,*

$$F\{|f - c| < \epsilon\} \leq C\epsilon^\gamma. \tag{2}$$

In the main result we will impose two further restrictions on f:

(F2) *f is uniformly continuous and $\int \|x\|^\beta f(x)dx < \infty$ for some $\beta > 0$.*

(F3) $\sup_{c\in[a,b]} F(\{f=c\}^h) = O(h)$, *where* $\{f=c\}^h = \bigcup_{x\in\{f=c\}} B(x,h)$ *and* $B(x,h)$ *denotes the closed ball in* $\mathbb{R}^d$ *with center* x *and radius* h.

Sufficient conditions for (F3) may be found in Walther (1997), Theorem 2. Essentially this condition requires the level sets of f to be smooth enough, without infinite peaks: a ball of positive radius should be able to roll along $\{f=c\}$.

We will define the smoothed quantile function as

$$V_n(\alpha) = \text{Leb}\{f_n > c_{n,\alpha}\} \quad \text{where } c_{n,\alpha} = \sup\{c : \int_{\{f_n>c\}} f_n > \alpha\} \quad \text{a.s.}$$

Choosing an adequate kernel K (bounded and without flat parts: $\text{Leb}\{K=c\}=0$ for every $c>0$), we will have $\tilde{F}_n\{f_n > c_{n,\alpha}\} = \alpha$ a.s. Thus we will consider kernels verifying

(K) K *is Lipschitz, without flat parts and has compact support with* $supp(K) \subseteq B(0,1)$.

If we assume that f is bounded and verifies (F1) then V is differentiable on $A = (\alpha-\delta, \alpha+\delta)$ and $v(\alpha)$, the derivative of V at α, is c_α^{-1} (see Polonik 1997). In this case we can define a smoothed version of the standardized quantile process defined in Einmahl and Mason (1992)

$$q_n(\alpha) = n^{1/2}\frac{V_n(\alpha)-V(\alpha)}{v(\alpha)}.$$

Our aim is to obtain convergence rates to 0 of

$$d_F(\{f > c_\alpha\}, \{f_n > c_{n,\alpha}\}) \tag{3}$$

as $n\to\infty$, where d_F is a pseudometric between sets defined as

$$d_F(C,D) = F(C\Delta D), \quad C, D \in \mathcal{B}_{\mathbb{R}^d}.$$

This measure-based distance is, together with the Hausdorff metric, the most frequently used in set estimation (see, for example, Korostelev and Tsybakov 1993). Molchanov (1998) studied the asymptotic behaviour of $\{f_n \leq c\}$ with respect to the Hausdorff metric, with f_n a generic estimator of f.

The convergence rates of (3) are interesting, for example, in the context of classifying an observation X coming from f as belonging to either the "core" of the distribution $\{f > c_\alpha\}$ or to the low probability region $\{f \leq c_\alpha\}$. If we estimate $\{f > c_\alpha\}$ by the plug-in estimator $\{f_n > c_{n,\alpha}\}$ and decide that $X \in \{f > c_\alpha\}$ when actually $f_n(X) > c_{n,\alpha}$, then (3) would represent the total probability of error of this procedure.

Baíllo, Cuesta-Albertos and Cuevas (2001) studied the asymptotic behaviour of the probability content of $\{f_n > c\}$ for fixed c. They also obtained rates of convergence of $F\{f_n > c_{n,\alpha}\}$ to α as a measure of the good performance of the plug-in estimator. However, it is not difficult to check that the coarse plug-in estimator of $\{f > c\}$ obtained from a histogram defined on a constant (not varying with n) partition attains $n^{-1/2}$ rates. This motivated the interest on a more adequate measure (such as (3)) of the performance of $\{f_n > c_{n,\alpha}\}$ as a set estimator of $\{f > c_\alpha\}$.

2. Convergence rates

In this section we will obtain rates for the a.s. convergence to 0 of (3). The first lemma decomposes the total probability of error (3) into a deterministic and a stochastic term and justifies the statement of condition (F1).

Lemma 1. *Assume that f is continuous, bounded by $M > 0$ and verifies condition*

(F1). Then, for every positive sequence $\epsilon_n \searrow 0$, there exists an n from which on

$$d_F(\{f > c_\alpha\}, \{f_n > c_{n,\alpha}\})$$
$$\leq C\epsilon_n^\gamma + M\epsilon_n^{-1}[n^{-1/2}q_n(\alpha) - (F\{f_n > c_{n,\alpha}\} - F\{f > c_\alpha\})] \quad a.s. \tag{4}$$

Proof of Lemma 1: Denote the set $\{f_n > c_{n,\alpha}\}\Delta\{f > c_\alpha\}$ by $D_n(\alpha)$. Then

$$d_F(\{f > c_\alpha\}, \{f_n > c_{n,\alpha}\}) \leq F\{|f - c_\alpha| \leq \epsilon_n\} + M\text{Leb}(D_n(\alpha) \cap \{|f - c_\alpha| > \epsilon_n\}).$$

We can bound the second term in the following way

$$\begin{aligned}
&\text{Leb}(D_n(\alpha) \cap \{|f - c_\alpha| > \epsilon_n\}) < \epsilon_n^{-1} \int_{D_n(\alpha)} |f - c_\alpha| \\
&= \epsilon_n^{-1}[\int_{\{f > c_\alpha\}} (f - c_\alpha) - \int_{\{f_n > c_{n,\alpha}\}} (f - c_\alpha)] \\
&= \epsilon_n^{-1}[c_\alpha(\text{Leb}\{f_n > c_{n,\alpha}\} - \text{Leb}\{f > c_\alpha\}) - (F\{f_n > c_{n,\alpha}\} - F\{f > c_\alpha\})].
\end{aligned}$$

□

Under condition (F1) we will consider the following classes of sets

$$\mathcal{C} = \{\{f > c\}, c \in [c_{\alpha+\delta}, c_{\alpha-\delta}]\} \quad \text{and} \quad \mathcal{C}_n = \{\{f_n > c\}, c \in [c_{\alpha+\delta}, c_{\alpha-\delta}]\},$$

where δ is a positive constant such that $c_{\alpha+\delta} > 0$ and $[c_{\alpha+\delta}, c_{\alpha-\delta}] \subset [a, b]$. Set

$$\bar{F}_n(t) = \sup\{\tilde{F}_n(C) : C \in \mathcal{C} \cup \mathcal{C}_n, \text{Leb}(C) \leq V(t)\}, \quad \alpha - \delta \leq t \leq \alpha + \delta.$$

Let

$$\bar{F}_n^{-1}(\alpha) = \inf\{t \in (\alpha - \delta, \alpha + \delta) : \bar{F}_n(t) > \alpha\}$$

be the generalized inverse of $\bar{F}_n$. Observe that, with probability 1, there exists an n_0 from which on $\bar{F}_n^{-1}(\alpha)$ is well defined. The following lemma focuses on the quantile process $q_n(\alpha)$ appearing in (4) and approximates $V_n(\alpha)$ through $V(\bar{F}_n^{-1}(\alpha))$.

Lemma 2. *Assume that f is bounded, continuous and verifies (F1). Assume also that K verifies (K). Then, under assumption (1),*

$$V_n(\alpha) = V(\bar{F}_n^{-1}(\alpha)) \quad a.s. \tag{5}$$

for n sufficiently large.

Proof of Lemma 2: From the definition of $\bar{F}_n^{-1}$ and the continuity of V it follows that, with probability 1, for n sufficiently large

$$\begin{aligned} V(\bar{F}_n^{-1}(\alpha)) &= \inf\{V(t) : \bar{F}_n(t) > \alpha,\ t \in (\alpha-\delta, \alpha+\delta)\} \\ &= \inf\left\{r \in (V(\alpha-\delta), V(\alpha+\delta)) : \alpha < \sup\{\tilde{F}_n(C) : C \in \mathcal{C} \cup \mathcal{C}_n, \mathrm{Leb}(C) \le r\}\right\}. \end{aligned}$$

Observe that under the hypotheses of the lemma $c_{n,\alpha} \to c_\alpha$ a.s. as $n \to \infty$ (see, for example, chapter 21 in van der Vaart 1998). Then, with probability one, for n sufficiently large $V_n(\alpha) = \inf\{\mathrm{Leb}(C) : C \in \mathcal{C} \cup \mathcal{C}_n, \tilde{F}_n(C) > \alpha\}$. So if we denote

$$\begin{aligned} S_1 &= \left\{r \in (V(\alpha-\delta), V(\alpha+\delta)) : \alpha < \sup\{\tilde{F}_n(C) : C \in \mathcal{C} \cup \mathcal{C}_n, \mathrm{Leb}(C) \le r\}\right\} \\ S_2 &= \{\mathrm{Leb}(C) : C \in \mathcal{C} \cup \mathcal{C}_n, \tilde{F}_n(C) > \alpha\}, \end{aligned}$$

we have that, for large n, $V(\bar{F}_n^{-1}(\alpha)) = \inf S_1$ and $V_n(\alpha) = \inf S_2$. If $r \in S_1$, then there exists a $C \in \mathcal{C} \cup \mathcal{C}_n$ with $\mathrm{Leb}(C) \le r$ and $\tilde{F}_n(C) > \alpha$. Then there is an $x \in S_2$ with $x \le r$. This implies that $V_n(\alpha) \le V(\bar{F}_n^{-1}(\alpha))$ a.s. On the other hand, if $r \in S_2$ there exists a $C \in \mathcal{C} \cup \mathcal{C}_n$ with $\mathrm{Leb}(C) = r$ and $\tilde{F}_n(C) > \alpha$. Hence $\alpha < \sup\{\tilde{F}_n(C) : C \in \mathcal{C} \cup \mathcal{C}_n, \mathrm{Leb}(C) \le r\}$ which yields $\inf S_1 \le r$. This implies $V_n(\alpha) \ge V(\bar{F}_n^{-1}(\alpha))$ a.s.

□

Theorem. *Assume that f is bounded and verifies (F1), (F2) and (F3). Assume also that the kernel K verifies (K). Then, if h is of exact order $(\log n/n)^{1/(d+2)}$,*

$$d_F(\{f > c_\alpha\}, \{f_n > c_{n,\alpha}\}) = O(n^{-\eta}) \quad a.s.$$

for any $0 < \eta < (d+2)^{-1}[\gamma/(1+\gamma)]$.

Proof: By Lemma 2 we know that

$$q_n(\alpha) = n^{1/2}\frac{v(\theta_n)}{v(\alpha)}[\bar{F}_n^{-1}(\alpha) - \alpha] \quad \text{a.s.},$$

for some θ_n between $\bar{F}_n^{-1}(\alpha)$ and α. Notice that, for n sufficiently large,

$$|\bar{F}_n^{-1}(\alpha) - \alpha| \leq \sup_{t\in[\alpha-\delta,\alpha+\delta]} |\bar{F}_n(t) - t| \quad \text{a.s.}$$

To see this, if $\bar{F}_n^{-1}(\alpha) \leq \alpha$ take a sequence $\{t_k\} \subset (\alpha-\delta, \alpha+\delta)$ with $t_k \searrow \bar{F}_n^{-1}(\alpha)$. Then, as $\bar{F}_n(t_k) > \alpha$, $|\bar{F}_n^{-1}(\alpha) - \alpha| = \alpha - \lim_{k\to\infty} t_k \leq \lim_{k\to\infty} |\bar{F}_n(t_k) - t_k| \leq \sup_t |\bar{F}_n(t) - t|$. If $\bar{F}_n^{-1}(\alpha) > \alpha$ take a sequence $\{r_k\}$ with $r_k \nearrow \bar{F}_n^{-1}(\alpha)$. Then, as $\bar{F}_n(r_k) \leq \alpha$, $|\bar{F}_n^{-1}(\alpha) - \alpha| = \lim_{k\to\infty}(r_k - \alpha) \leq \lim_{k\to\infty} r_k - \bar{F}_n(r_k) \leq \sup_t |\bar{F}_n(t) - t|$.

But for each $t \in [\alpha - \delta, \alpha + \delta]$

$$\begin{aligned}
\bar{F}_n(t) - t \quad &= \sup\{(\tilde{F}_n(C) - F\{f > c_t\}) : C \in \mathcal{C} \cup \mathcal{C}_n, \mathrm{Leb}(C) \leq V(t)\} \\
&\leq \sup\{(\tilde{F}_n - F)(C) : C \in \mathcal{C} \cup \mathcal{C}_n, \mathrm{Leb}(C) \leq V(t)\} \\
&\leq \sup_{c\in[c_{\alpha+\delta}, c_{\alpha-\delta}]} (\tilde{F}_n - F)\{f_n > c\} + \sup_{c\in[c_{\alpha+\delta}, c_{\alpha-\delta}]} (\tilde{F}_n - F)\{f > c\}. \qquad (6)
\end{aligned}$$

This, Lemma 1 and the fact that $F\{f > c_\alpha\} = \tilde{F}_n\{f_n > c_{n,\alpha}\} = \alpha$ a.s. means that we can obtain rates for $d_F(\{f > c_\alpha\}, \{f_n > c_{n,\alpha}\})$ as a by-product of rates for the suprema appearing in (6).

Concerning the supremum of $(\tilde{F}_n - F)\{f_n > c\}$, fix $\Delta > 0$ such that $[c_1, c_2] \subset [a, b]$, where $c_1 = c_{\alpha+\delta} - \Delta$ and $c_2 = c_{\alpha-\delta} + \Delta$. Observe that, under the hypotheses of the theorem, $\sup_x |f_n(x) - f(x)| \to 0$ a.s. (see Prakasa Rao 1983). This implies that there exists an n from which on $\sup_x |f_n(x) - f(x)| < \Delta$. In particular $\{f > c_1\} \subset \{f_n > c\} \subset \{f > c_2\}$ a.s. for any $c \in [c_{\alpha+\delta}, c_{\alpha-\delta}]$. As a consequence, for some constant $C > 0$,

$$\int_{\{f_n>c\}} (f_n - f) \quad \leq \int (f_n - f) 1_{\{f>c_1\}} + \int (f_n - f) 1_{\{f_n \geq f, c_2 \geq f > c_1\}}$$

$$\leq \sup_{c\in[c_1,c_2]} \int_{\{f>c\}} (f_n - f) + C\left(\frac{\log n}{nh^d}\right)^{1/2},$$

where we have used a result by Stute (1984) concerning the convergence rate to 0 of $\sup|f_n(x) - f(x)|$ over compact sets.

Next we will study $\sup_{c\in[c_1,c_2]}(\tilde{F}_n - F)\{f > c\}$. Observe that

$$\int_{\{f>c\}} (f_n - f) = (F_n - F)\int 1_{\{f(\cdot+y)>c\}} K_h(y)dy +$$
$$+ \int\int [1_{\{f(x+y)>c, f(x)\leq c\}} - 1_{\{f(x+y)\leq c, f(x)>c\}}] dF(x)K_h(y)dy$$

Thus

$$\sup_{c\in[c_1,c_2]} \int_{\{f>c\}} (f_n - f) \leq \sup_{c\in[c_1,c_2]} (F_n - F)\int 1_{\{f(\cdot+y)>c\}} K_h(y)dy + 2\sup_{c\in[c_1,c_2]} F(\{f=c\}^h).$$

By assumption the second term in the last expression is $O(h)$. Concerning the first term, by McDiarmid's inequality (McDiarmid 1989), it is only necessary to check the rates to 0 of

$$E\sup_{c\in[c_1,c_2]} (F_n - F)\int 1_{\{f(\cdot+y)>c\}} K_h(y)dy,$$

which, for n large enough, is smaller than $E\sup_{c\in[c_1,c_2]}(F_n - F)\{f > c/2\}$, as f is uniformly continuous and $\text{supp}(K) \subseteq B(0,1)$. By Vapnik-Chervonenkis inequality the last expectation is $O(n^{-1/2})$ (see for example Devroye, Györfi and Lugosi 1996). We have finally obtained that,

$$d_F(\{f > c_\alpha\}, \{f_n > c_{n,\alpha}\}) \lesssim \epsilon_n^\gamma + \epsilon_n^{-1}[h + n^{-1/2} + (\log n/(nh^d))^{1/2}] \quad \text{a.s.},$$

where $\lesssim$ denotes "smaller than up to a constant". This yields the desired result.

□

Remark: As we see, the plug-in approach relates the asymptotic properties of level set estimators to the behaviour of density f in a neighbourhood of c_α, in particular

to the parameter γ (see also Tsybakov 1997). The result in the theorem confirms the intuitive idea that, the steeper f is around c_α (large γ), the faster the rates will be.

Acknowledgements: I would like to thank Antonio Cuevas for his suggestions, which led to improvement of the results.

REFERENCES

Baíllo, A., Cuesta-Albertos, J. A. and Cuevas, A. (2001). Convergence rates in nonparametric estimation of level sets. *Stat. Prob. Letters* 53, 27-35.

Cuevas, A., Febrero, M. and Fraiman, R. (2000). Estimating the number of clusters. *Canad. J. Statist.*, 28, 2, 367-382.

Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Springer: New York.

Devroye, L. and Wise, G. (1980). Detection of abnormal behaviour via nonparametric estimation of the support. *SIAM J. Appl. Math.* 38, 480-488.

Einmahl, and Mason, D. (1992). Generalized quantile processes. *Ann. Stat.* 20, 2, 1062-1078.

Hartigan, J. A. (1975). *Clustering Algorithms.* New york: Wiley.

Korostelev, A. P. and Tsybakov, A. B. (1993). *Minimax Theory of Image Reconstruction.* New York: Springer-Verlag.

McDiarmid, C. (1989). On the method of bounded differences. *London Mathematical Society Lecture Notes Series* 141, 148-188. Cambridge: Cambridge University

Press.

Molchanov, I. S. (1998). A limit theorem for solutions of inequalities. *Scand. J. Statist.* , 25, 235-242.

Polonik, W. (1995)). Measuring Mass Concentrations and Estimating Density Contours Clusters-an Excess Mass Approach. *Ann. Statist.*, 23, 3, 855-881.

Polonik, W. (1997). Minimum volume sets and generalized quantile processes. *Stoch. Proc. Appl.* 69, 1-24.

Prakasa Rao, B. L. S. (1983). *Nonparametric Functional Estimation.* Academic Press: New York.

Stute, W. (1984). The oscillation behavior of empirical processes: the multivariate case. *Ann. Prob.* 12, 2, 361-379.

Tsybakov, A. B. (1997), On nonparametric estimation of density level sets. *Ann. Statist.* 25, 3, 948-969.

van der Vaart (1998). *Asymptotic Statistics.* Cambridge: Cambridge University Press.

Walther, G. (1997). Granulometric smoothing. *Ann. Statist.*, 25, 6, 2273-2299.